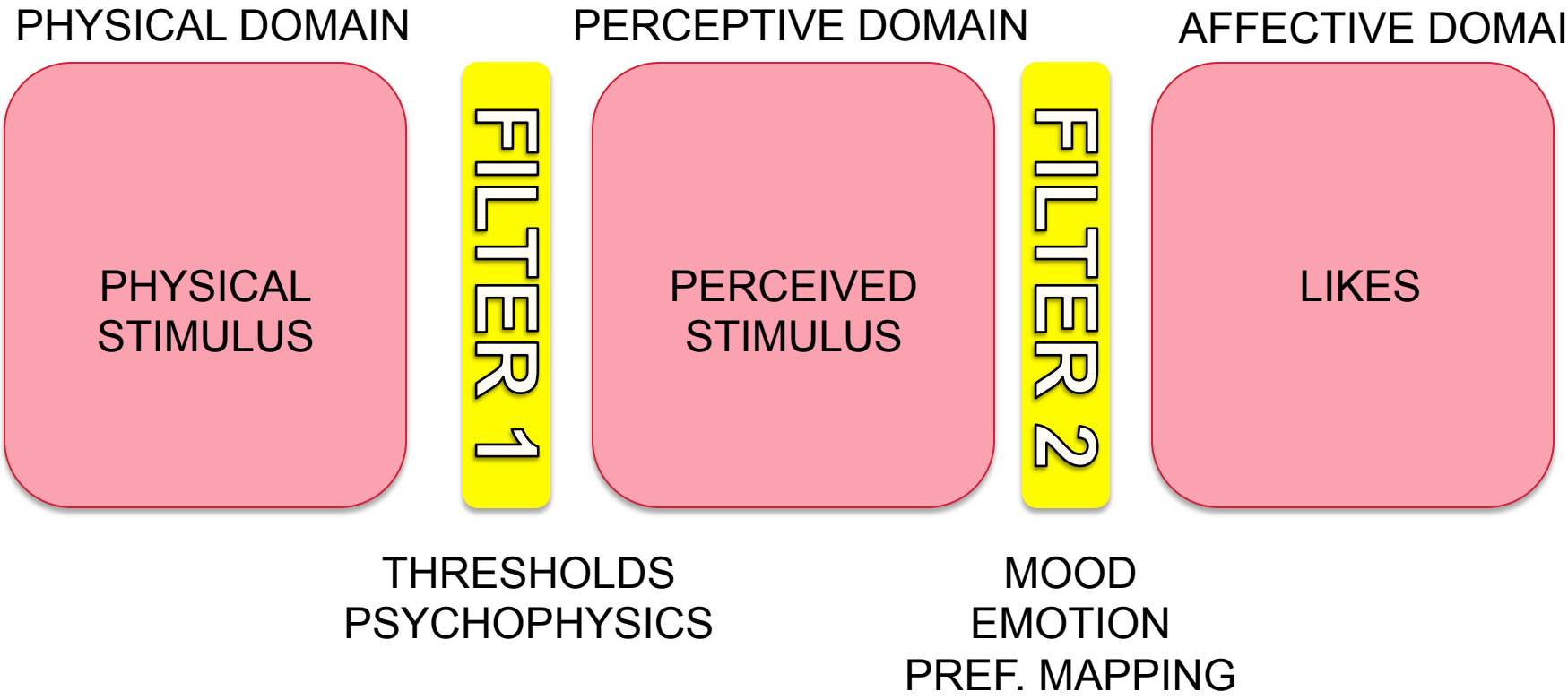
How to Design and Conduct Listening Tests for Audio and Acoustics

Bruno Fazenda

School of Computing, Science and Engineering
Acoustics Research Centre

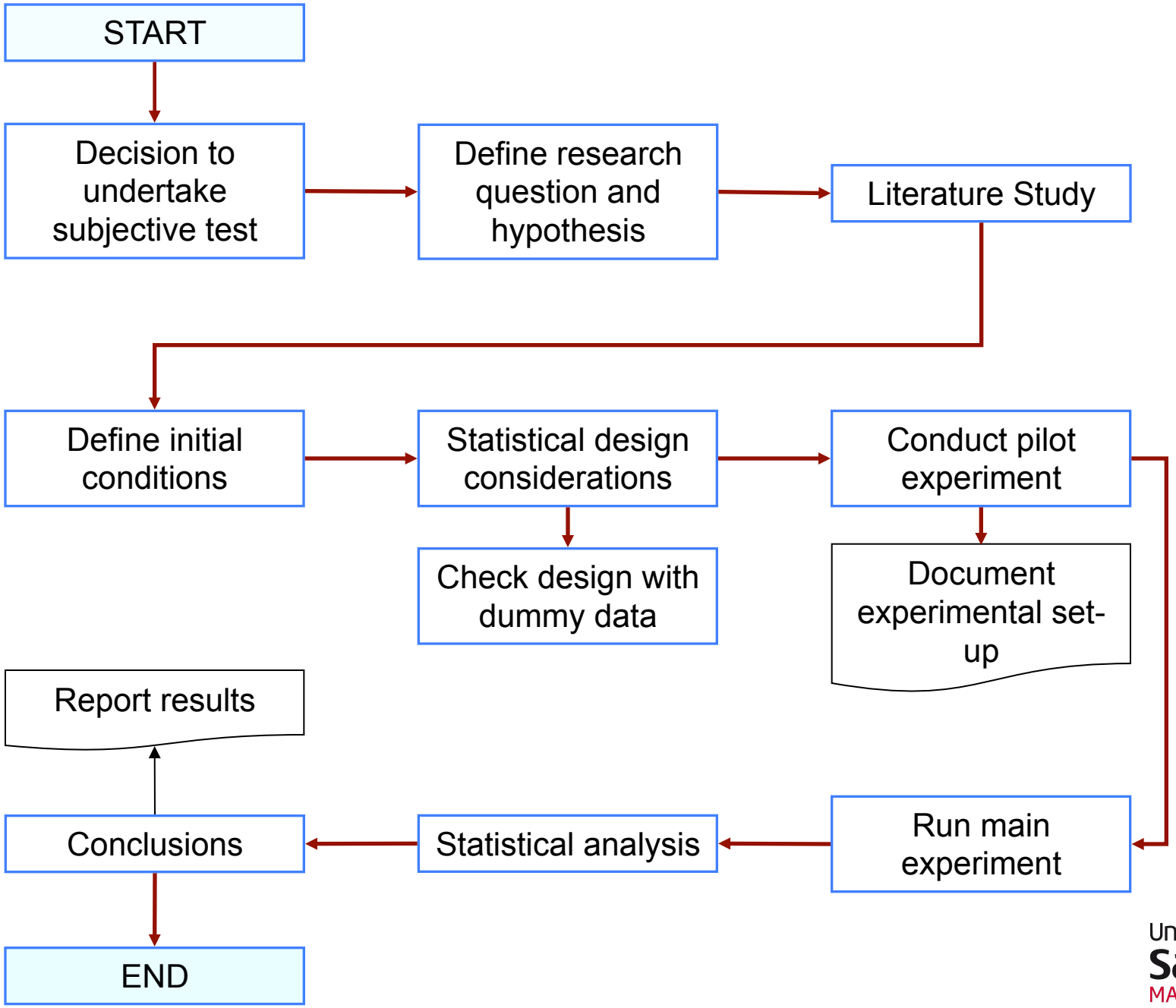
University of Salford MANCHESTER

What is subjective testing?



Bech, S., Presented at S3A Sandpit July 2014

University of Salford MANCHESTER



University of Salford MANCHESTER

Introduction

Subjective testing is a lengthy process which requires preparation and knowledge in order to provide credible results

When to do listening tests?

- Find preference of a system/method over others
- Identify problems introduced by a system/method
- Quantify subject response to a given stimuli

When NOT to do listening tests?

- If a similar test has been done
- If there is an (automatic) objective way of doing it
  - i.e. measure using a model/method/apparatus

The research question and hypothesis

- What is it that we are trying to test?
- Example 1:
  - Question: How does the sound quality of our loudspeaker A compare with that of loudspeaker B?
  - Null Hypothesis(H_0): The perceived sound quality of loudspeaker A is identical to loudspeaker B.
- Example 2:
  - Question: Do ambisonic systems provide better localisation cues than surround systems?
  - Null Hypothesis(H_0): Localisation in ambisonics is identical to surround.
- The primary aim of the test is to collect data and check whether it rejects the Null Hypothesis

The importance of a good literature review

What has been written about this previously?

- Journals, books, website

Is there an established method?

- If so, use it!
- Otherwise you can ‘adapt’ a method used previously for a similar test

Check usual publications in field

- JAES, JASA, ITU

Initial Conditions

The initial condition is defined by a set of variables that will test the hypothesis

Variables can be:

- Categorical – eg: Age (in ranges 18-25, 26-30…), Loudspeaker(A,B,C,D)…
- Numerical – eg: cutoff frequency, SPL…

Test design, deployment and statistical analysis techniques will be dependent on these

Variables in a statistical analysis

- Dependent – the answer provided by the subject
- Independent – those controlled by the experimenter

Examples:

Examples of Test Variables

Test	Independent Variables	Dependent Variables
Example 1: Loudspeaker comparison	▪Loudspeakers A and B ▪Musical Genre/Samples ▪Loudspeaker position in room ▪Presentation Levels	▪Subject's preference in the form of ranking or rating scores
Example 2: Ambisonic vs Surround	▪Ambisonic and surround systems ▪Musical Genre/Samples ▪Transient/Frequency characteristics of signals (eg: snare hits vs bowed cello)	▪Subject's localisation as perceived angle

Statistical Design Considerations

Who is going to take my test?

- How do I find subjects?
- Do they need to conform to given prerequisites, ie: Males aged 25-35

How many subjects are needed for statistical significance? (more on significance later)

- Chicken and egg problem
- Small perceptual differences or 'very subjective' tasks - large number of subjects (>30) to ensure 'normal distribution' and stronger parametric tests
- *Statistical Power* - number of required samples (i.e. subjects x nr auditions) for statistical significance
  - Pre-hoc: this usually requires an estimation of parameters in data such as the variance
  - Post-hoc: will tell you whether a statistical non-significant is due to low power or not having an effect from the experiment

Will we use experts or naïve listeners*?

Healthy hearing

Olive, Sean E. "Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: a case study." *Journal of the Audio Engineering Society* 51.9 (2003): 806-825.

Experimental Setup

Where will I conduct my test?

- Usually need quiet, standardised and controlled conditions
  - Background and equipment noise
  - Room effects
    - Listening room, anechoic, studio, headphones

Equipment

- Calibration
- Levels

Everything needs to be properly documented!

Example Analysis

- We tested two loudspeakers (A: B&W and B: Behringer)
- On day 1 we played a Dub/Reggae music sample through A to all subjects
- On day 2 we played Jazz through B to all subjects
- Results: All subjects considered A to be better than B
- ?

Analysis:

- Independent variables?
  - Loudspeakers (A and B); Music Samples (Dub/Reggae and Jazz)
  - Time (Day 1 and Day 2)
- Problems?
  - Confounding factors – Order Bias; Music sample bias
  - Preconceived ideas (branding)

Example Analysis

Re-Design:

- Single listening session (same day) if possible
- Acoustic curtain; blindfolding
- Full combinations (*Full factored analysis*) of all independent variables
  - Every subject listens to both music samples through both speakers
- Randomisation of Independent Variables

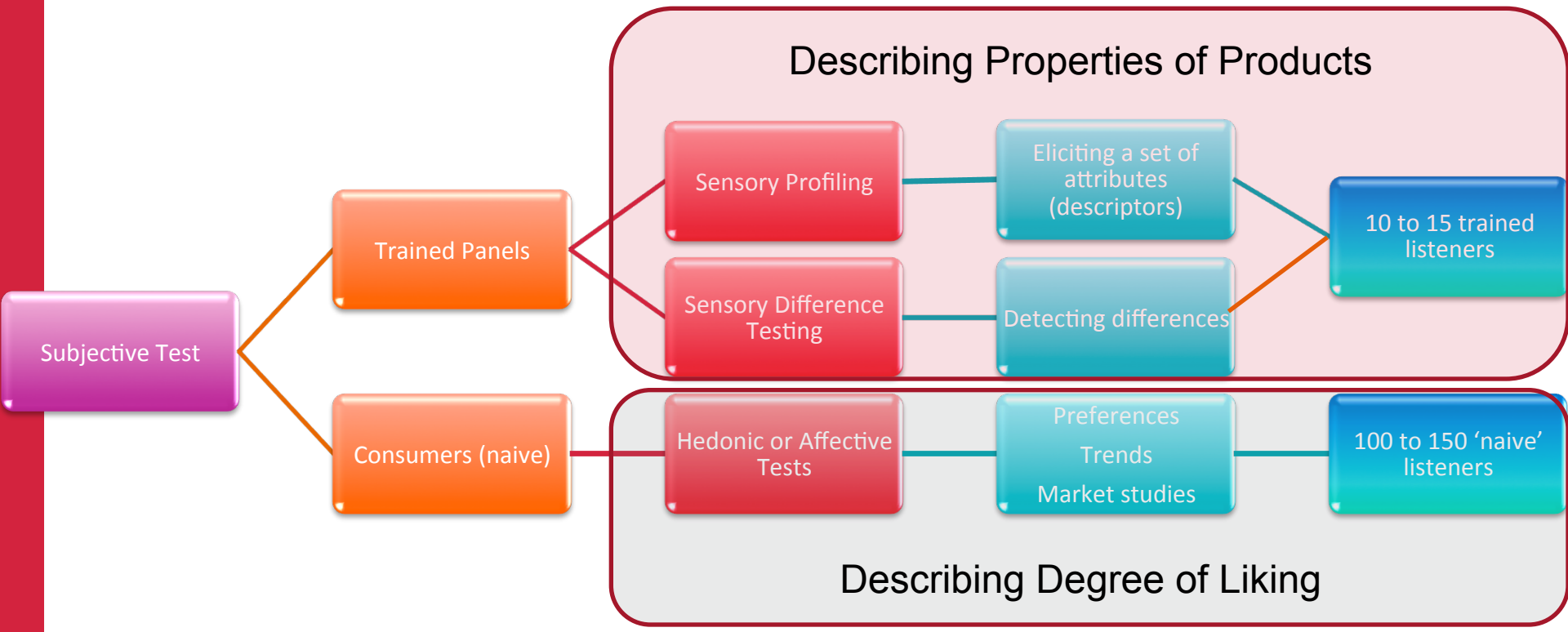
A pilot experiment highlights:

- Any practical problems
- Suitability of the samples being used
- The amount of time required to perform the test

Choosing the correct method

DEFINING THE TEST

Defining the Test - Subjects



Sensory Panel Methodologies (Example)

Method:

1) **Selection of Subjects**
   a. A list of Salford subjects - hearing should have been tested
   b. Short questionnaire
   c. Discrimination and first word elicitation phase - individual

2) **DA Panel Session 1**
   a. Arrange a group discussion to reduce terms / agree on their meaning

3) **DA Panel Session 2**
   a. Another group discussion
   b. Finalise terms
   c. Decide on end points and the scales
   d. Write descriptions

4) **Testing**
   a. DA Panel to rate samples according to new descriptors
   b. Naive Panel to rate in a simple quality exercise
   c. Naive Panel to rate samples according to new descriptors

5) **Analysis**
   a. Principle Component Analysis
   b. Multiple dependant and independent variables so we can link the two with stats

6) Discussion
   a. Any relationships found
   b. The correlation between DA and Naive listeners
   c. Usefulness of these descriptors in further testing

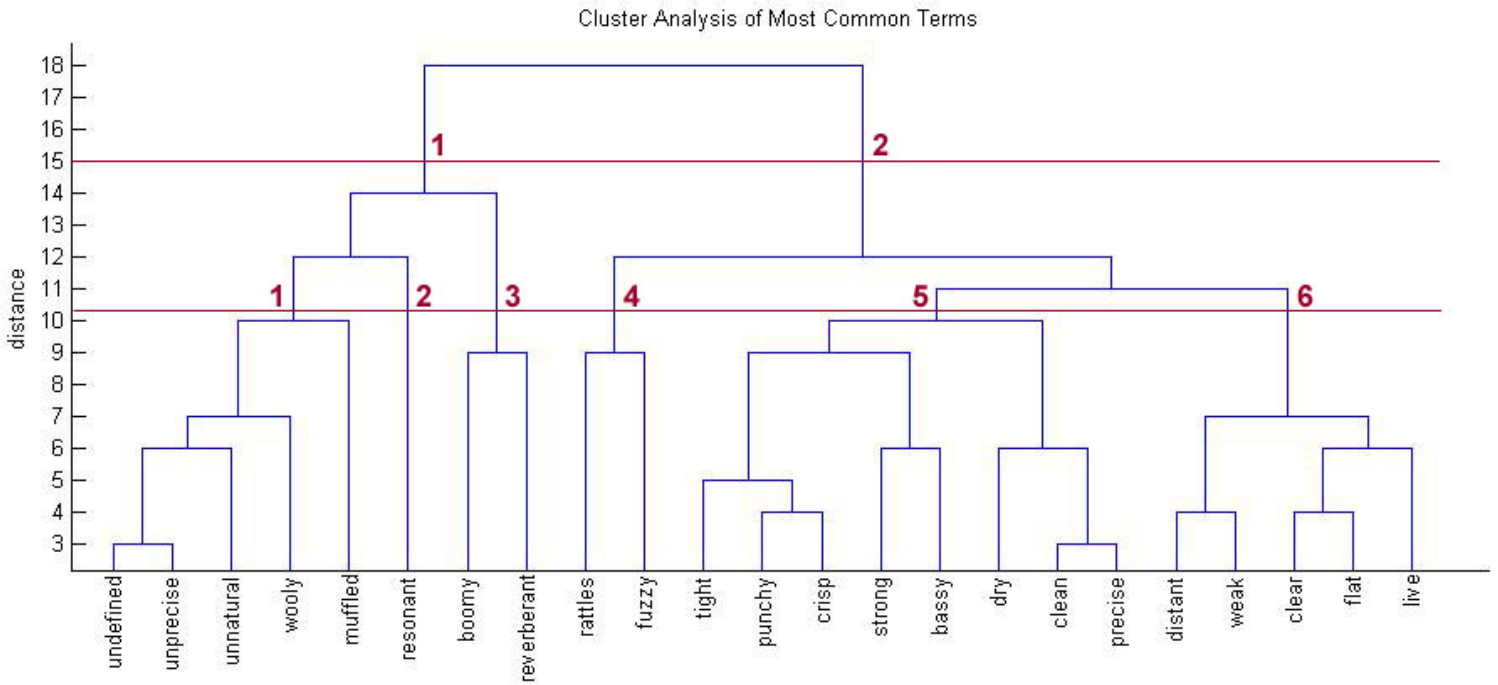
Implementation:

1) What samples am I going to use? How will I create them? Linked to my room model? There are pros and cons to linking them and not.
2) Write software to:
   a. Test ability to discriminate samples
   b. Elicit words originally from individual subjects
   c. Test samples on a single scale
   d. Test samples on multiple scales (one for each descriptor)
3) Book studios / subjects
4) Try and understand the stats!

Sensory Panel Methodologies (Example)

Descriptive Sensory Analysis (sensory Profiling)

- Identifying product properties through a consensus language
- Eliciting (creating) attributes to use for describing product differences
- Product examples are used to elicit descriptors

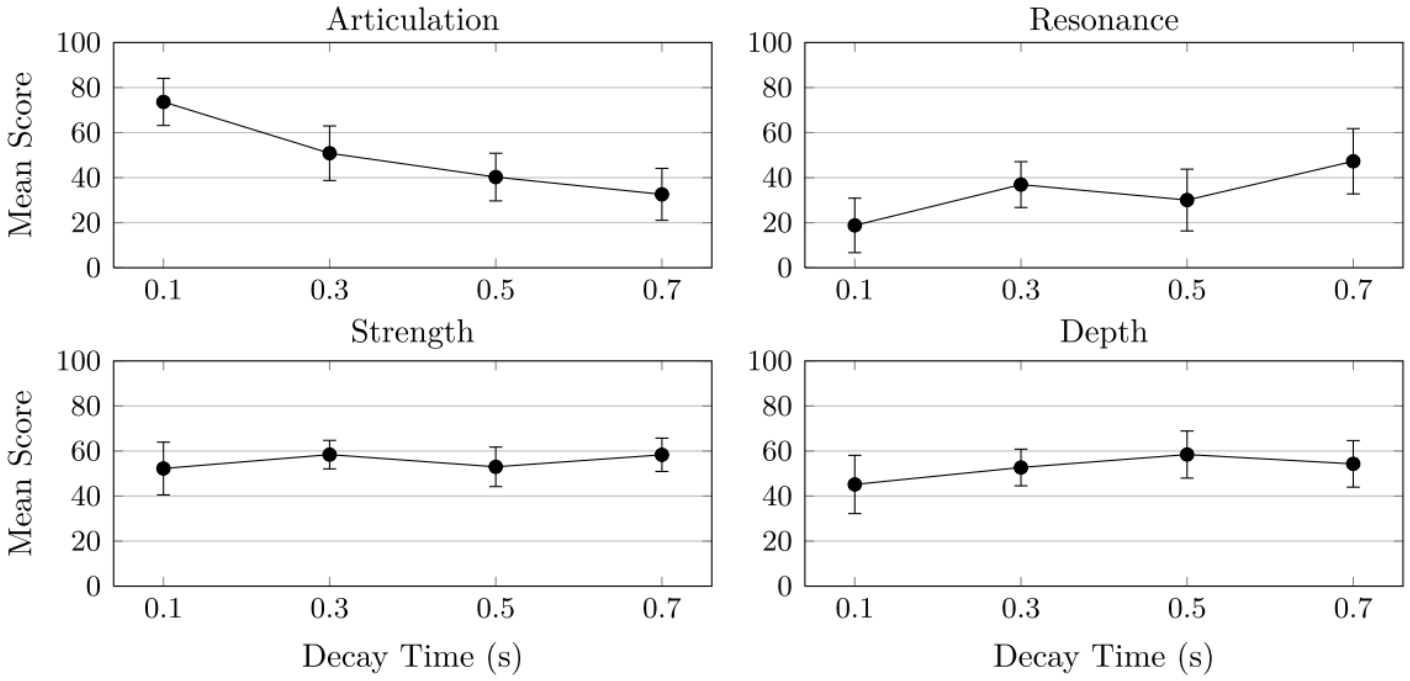


University of Salford MANCHESTER

Sensory Panel Methodologies (Example)

Descriptive Sensory Analysis (sensory Profiling)

- Scales may also be drawn with consensual end points (i.e. ‘strong’ vs ‘weak’ for bass strength in a mix)
- New tests can now be carried out

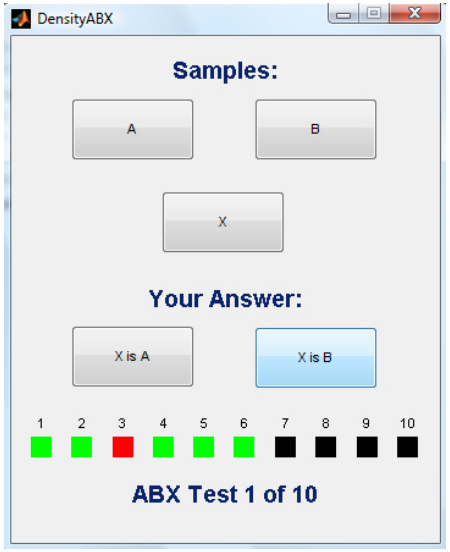


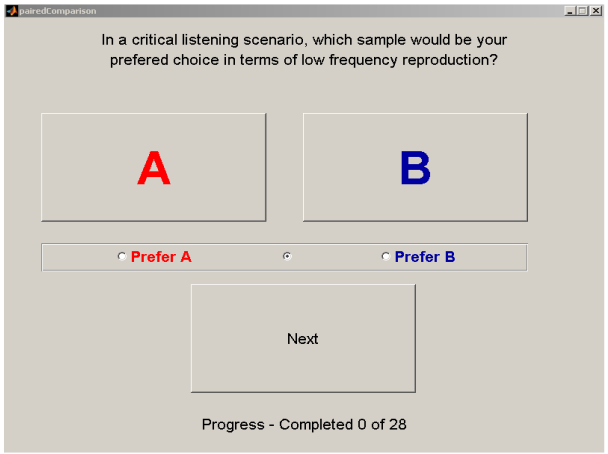
Change in decay time - mean scores and 95% confidence intervals

University of Salford MANCHESTER

Discrimination Tests

- Can users detect small differences between products?
- Triangle test
  - Two identical and one different sample
  - Which one is different?
- ABX
  - A double blind test methodology
  - Sample X is a random choice between A and B
  - Each subject runs test for a number of trials (>10)
  - Number of trials gives statistical significance based on binomial distribution
- Paired comparison (Thurstone)
  - Which of the samples has/is more <attribute> ?





University of Salford MANCHESTER

Consumer Tests

Rating Based

- Consumer provides a score for each product on a given scale

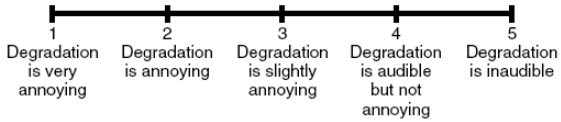
Ranking Based

- All samples (products) are presented simultaneously
- Consumers asked to rank them in order of preference
- Can use (Thurstone's) paired comparison but this will lead to a large number of auditions

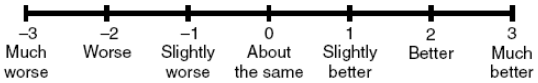
Rating Based Scales

- Mainly used for perceived quality
  - "Rate the quality of product X"
- Also used for noticeability of degradations
  - See examples

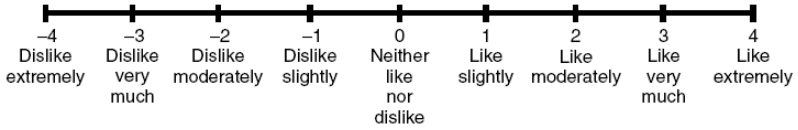
(a) Absolute category rating (ACR) scale [224]



(b) Degradation category rating (DCR) scale [224]



(c) Comparison category rating (CCR) scale [224]



(d) 9-point hedonic categorical scale [342]

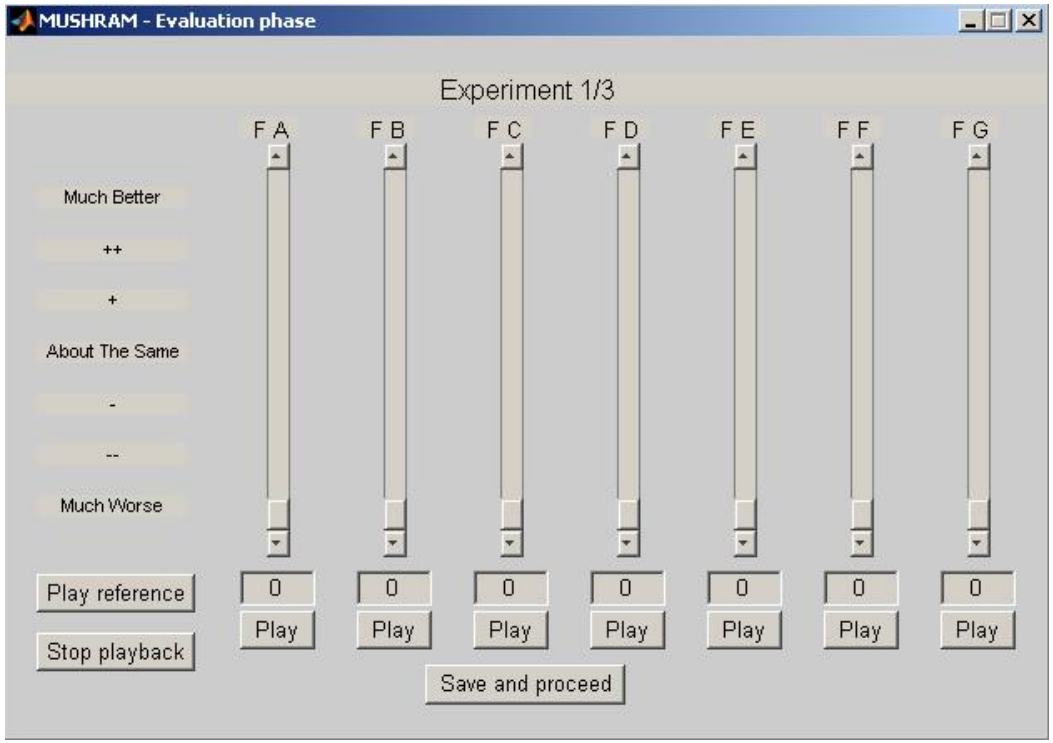
(e) ITU-R 5-point continuous impairment scale [203]

(f) ITU-T P.1534 continuous scale [217]

From: Bech, Zacharov, *Perceptual Audio Evaluation,* Wiley, 2006

Rating Based Scales

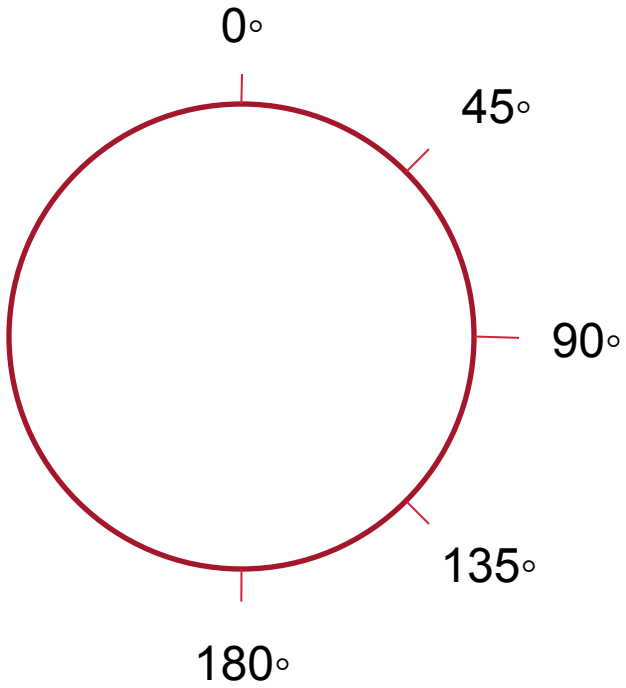
- A common example in audio is the **MU**ltiple **S**timuli with **H**idden **R**eference and **A**nchor – MUSHRA
  - ITU-R recommendation BS.1534-1



From: Bech, Zacharov, *Perceptual Audio Evaluation,* Wiley, 2006

Numerical Continuous Scale

- Numerical continuous scale
- Eg: ‘After listening to the sample please indicate perceived direction in degrees (use diagram below as a guide)’



University of Salford MANCHESTER

Documenting a Test

- What should be recorded?
  - Equipment details
  - Calibration
  - Nr Subjects
  - Physical Setup – positions, heights, etc
  - Sample names/numbers;
  - Full test procedure:
    - Subject sits on chair
    - Given consent form and instructions
    - Definitions
  - Health and safety
  - Photos
  - etc

Statistics

Analysing your results using statistical methods is the only way to prove experimental outcomes

Some people are usually scared of ‘stats’

- However, there are some tools (Matlab, Excel, SPSS) that will allow you to input the data and get the results
- But you NEED to know how to use the tool and what the results mean

Statistics

- You can't test every single user of your technology/sample/ theory, you only get an estimate of the outcomes
- Therefore, a statistical test is required to find the probability that the outcomes obtained during the experiment were found by chance
  - One of the results from the statistical test is *p* (or significance) value
- Which is compared to a *critical value* α set by the experimenter
  - For listening tests α = 0.05 (or 5% probability)
- The statistical test effectively identifies whether the null hypothesis can be rejected
- So, if $p < \alpha$ we usually have evidence to reject the null hypothesis

Worked Example

H_0: “We are unable to hear a difference between loudspeaker cables A and B”

Independent variables:

- Cables (x2)
- Musical Sample (x2)

6 subjects

- 10 auditions each

ABX test

Results:

Nr of successful identifications out of 10

	Music Sample 1	Music Sample 2
John	10	6
Tom	9	7
Hannah	9	4
Matt	8	5
Jenny	10	6
Peter	7	6
total observed	53	34
total n	60	60

How can we reliably tell whether the subjects detect a difference between cables A and B?

A Worked Example

ABX test

Common Statistical analysis for this is the binomial test – check following example

We want to find out whether subjects are reliably telling the difference between the cables or guessing (usually set at 50% correct answers)

Using binomial test

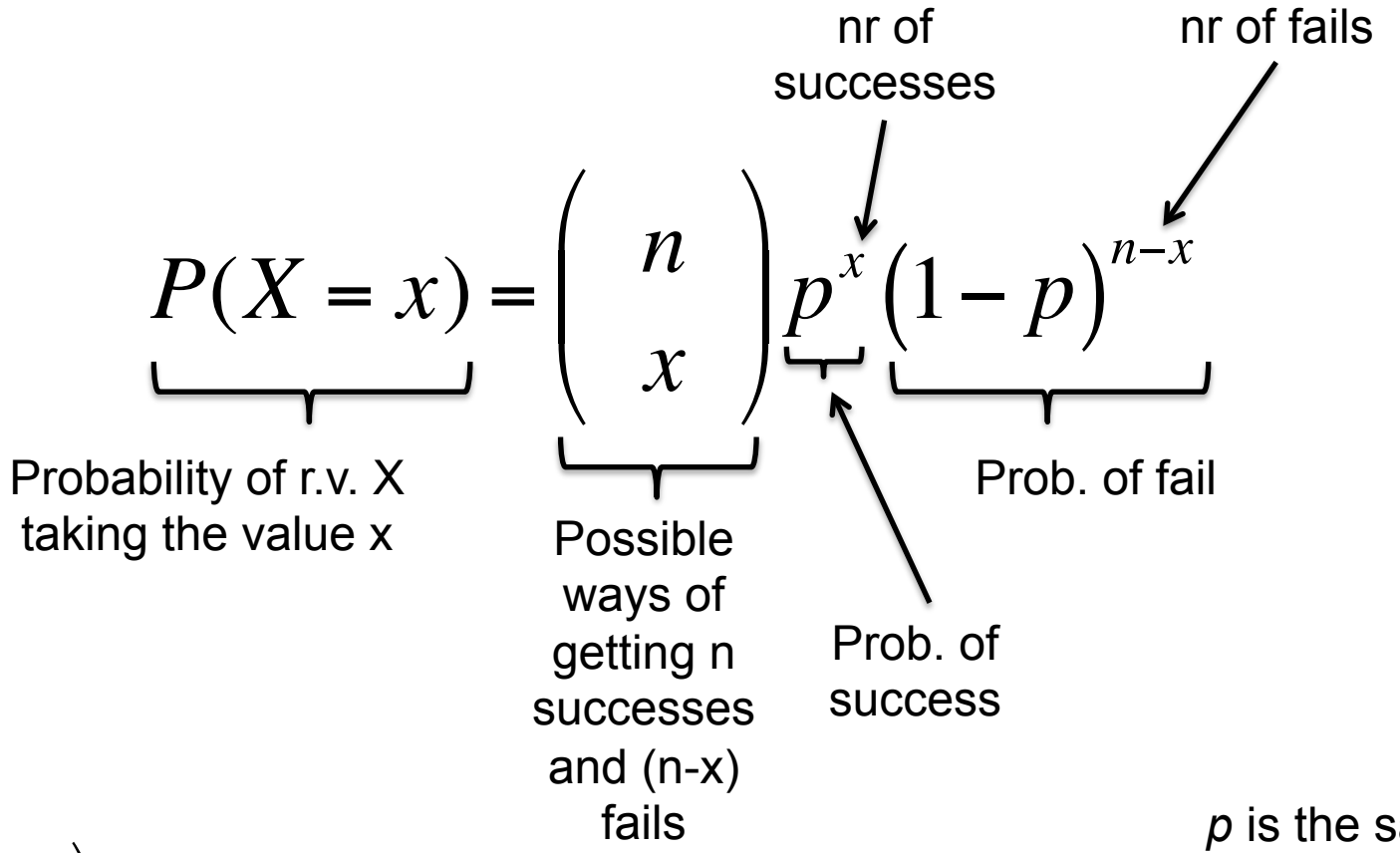
The test is based on *bernoulli* trials

- A test where the outcome is one of two possible
  - Success or fail
- The *probable* number of successes *X* in *n* bernoulli trials can be found from the *binomial distribution*

The probability is worked out from the following:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Using binomial test



$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

p is the sample proportion, which in our case is set at 0.5 or 50% chance rate.

University of Salford MANCHESTER

A Worked Example

In an ABX test we are interested in finding evidence that subject(s) *can* tell a difference between the two cables

- i.e. We want to know the probability of getting x successes in n trials

For this we need to start by setting out *hypothesis*

- H_0: Subject detection is at guess rate (50%)
- H_1: Subject detection is above guess rate (50%)

In our test, if we find evidence to reject the H_0 we can say that subject can reliably tell the difference between cables A and B

A Worked Example

Let's work this out for a single subject (Peter) with sample 1

- Remember
  - H_0: Subject detection is at guess rate (i.e. p=0.5)

Nr of successful identifications out of 10

	Music Sample 1	Music Sample 2
John	10	6
Tom	9	7
Hannah	9	4
Matt	8	5
Jenny	10	6
Peter	7	6
total observed	53	34
total n	60	60

$$p(X=7)=\binom{10}{7}0.5^{7}\left(1-0.5\right)^{10-7}=0.1171$$

– The probability that Peter is guessing is 12%
  - Which is above the set critical level of 5%
  - So we retain H_0

REMINDER: $P(X=x)=\binom{n}{x}p^{x}\left(1-p\right)^{n-x}$

University of Salford MANCHESTER

A Worked Example

Nr of successful identifications out of 10

	Music Sample 1	Music Sample 2
John	10	6
Tom	9	7
Hannah	9	4
Matt	8	5
Jenny	10	6
Peter	7	6
total observed	53	34
total n	60	60

What about Hannah?

- Remember
  - H_0: Subject detection is at guess rate (i.e. p=0.5)

$$p(X=9)=\binom{10}{9}0.5^9\left(1-0.5\right)^{10-9}=0.0097$$

– The probability that Hannah is guessing is below 1%
  - Which is well below the set critical level of 5%
  - So we reject H_0, Hannah was not guessing

REMINDER: $P(X=x)=\binom{n}{x}p^x\left(1-p\right)^{n-x}$

University of Salford MANCHESTER

A Worked Example

In most cases we want to look at results from all subjects

- i.e.

Nr of successful identifications out of 10

	Music Sample 1	Music Sample 2
John	10	6
Tom	9	7
Hannah	9	4
Matt	8	5
Jenny	10	6
Peter	7	6
total observed	53	34
total n	60	60

$$P(X=53)=\binom{60}{53}0.5^{53}(1-0.5)^{60-53}=3.35e-3$$

– The probability that the group is guessing is very small
  - And below the set critical level of 5%
  - So we reject $H_{0,}$ There IS a perceived difference between cables

REMINDER: $P(X=x)=\binom{n}{x}p^{x}(1-p)^{n-x}$

University of Salford MANCHESTER

A Worked Example

What about using music sample 2?

Nr of successful identifications out of 10

	Music Sample 1	Music Sample 2
John	10	6
Tom	9	7
Hannah	9	4
Matt	8	5
Jenny	10	6
Peter	7	6
total observed	53	34
total n	60	60

$$P(X=34)=\binom{60}{34}0.5^{34}(1-0.5)^{60-34}=0.06$$

- The probability that the group is guessing is 6%
  - Which is above the set critical level of 5%
  - So we retain H_0 for music sample 2:
  - Using music sample 2 subjects cannot detect differences between speaker cables

REMINDER: $P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$

University of Salford MANCHESTER

Reporting the Results

Conclusions from Stats Analysis:

- Some subjects were reliable in identifying differences between cables ($P<0.05$) and some weren't ($p>0.05$)
- The group could reliably identify differences when using Music sample 1 ($p<0.05$)
- The group could **not** reliably identify differences when using Music sample 2 ($p>0.05$)

This means:

- There is a significant quality difference between cables (mock data has been presented here!)
- Music sample 1 seems to be better at revealing these differences

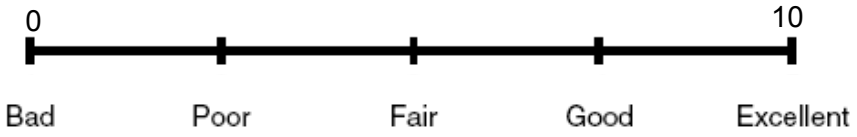
Testing various levels – a worked example

What if we want to test something at various levels?

Eg: We want to rate 4 loudspeakers in terms of perceived quality?

Testing for Quality – a worked example

- First we need to **define quality** so our subjects know what they are evaluating
  - "Result of an assessment of the perceived auditory nature of a sound with respect to its desired nature."*
- Then we need to decide the scale to be used
  - Example adapted from ACR in ITU



(a) Absolute category rating (ACR) scale [224]

*Jekosch, U. (2004). Basic Concepts and Terms of. acta acustica united with Acustica, 90(6), 999-1006.

University of Salford MANCHESTER

Testing various levels – a worked example

Rating scores between 0 and 10 in steps of 0.1

6 listeners

Results:

Loudspeakers A, B, C, D

	A	B	C	D
John	4.5	9.6	3.1	10
Tom	6.7	9.4	2.9	9.6
Hannah	3.2	9.6	2.2	9.1
Matt	6.6	9.2	3.4	9.4
Jenny	5.3	8.7	3.2	8.2
Peter	5.2	9.3	2.8	8.9

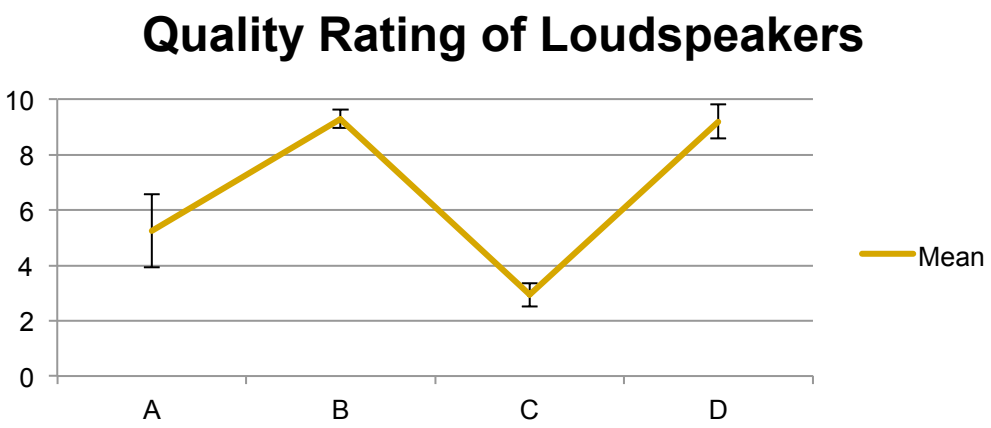
Further Stats Analysis

Usually we are looking for the *mean* scores

- Standard deviation or any other measure of *variance* is also quite useful as it tells you the agreement between subjects

Loudspeakers A, B, C, D

	A	B	C	D
John	4.5	9.6	3.1	10
Tom	6.7	9.4	2.9	9.6
Hannah	3.2	9.6	2.2	9.1
Matt	6.6	9.2	3.4	9.4
Jenny	5.3	8.7	3.2	8.2
Peter	5.2	9.3	2.8	8.9
Mean	5.2500	9.3000	2.9333	9.2000
St. Dev	1.32	0.33	0.42	0.62



From this data we can make some inferences

- B and D are scored similarly
- C is scored worse

But we need to prove this statistically!

University of Salford MANCHESTER

Analysis Of VAriance

The ANOVA checks the hypothesis that *all means are the same* (for the 4 loudspeakers)

Can be run in excel (or Matlab, SPSS, R, etc)

It gives a percentage probability that they all belong to the same *population*

This is output in *P-value* in the table

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	175.7813	3	58.59375	97.1033	4.3E-12	3.098391
Within Groups	12.06833	20	0.603417			
Total	187.8496	23				

If this value is below 0.05 (5%) then you have significant evidence that the means are NOT the same

- i.e. you have a statistically significant result:
  - The loudspeakers have different qualities according to your panel of listeners

That is the case in the data shown in Excel file and table above

It tells you that there is a difference between **at least** two of the mean scores

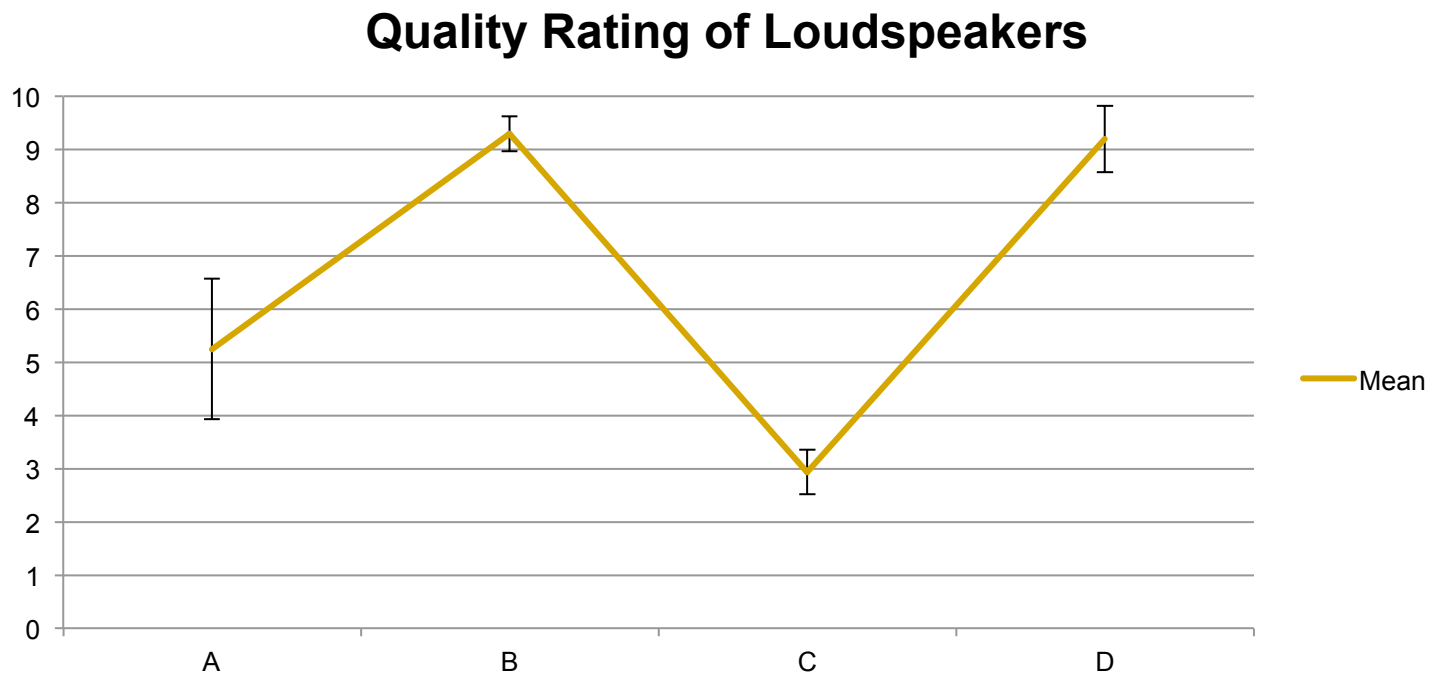
University of Salford MANCHESTER

Taking testing further

The ANOVA has suggested a *significant effect* for loudspeakers

But it *only* tells you that at least one loudspeaker is scored significantly different from the others

But how do we tell if there is a difference, say, between B and D?



University of Salford MANCHESTER

The paired comparison t-test

If you want to find out which particular scores are significantly different you would need to run a series of paired comparisons

I.e. you run a stats analysis for each paired combination in your test

One option is the t-test (but this needs to be corrected!)

SPSS does this with Bonferroni corrections

Bit of stats reading for you…

ANOVA Variations

In our loudspeaker example, you have a single factor (loudspeakers)

- So you have done a one way anova

If you collect results for N independent variables (eg: Loudspeaker, Music Samples, Room, Level, etc) you would need a N-way anova

- This would give you a *p* value for each independent variable but also, and more importantly, a *p* value for the interactions between them
- Checking for interactions (mostly 2nd order ones) is a very valuable tool to extract more meaningful data from your test

Tests with and without replication

- If each listener participates in all tests you would have a N-factor *with replication*
- If different groups of listeners evaluate different levels of a variable then it would be a N-factor *without replication*

Further Statistical Evidence

You can further support your results by calculating:

- Observed Statistical Power
  - This was mentioned when deciding on number of subjects/ sampling
  - It gives you an indication whether a non-significant result was obtained either due to low sampling or a ‘real’ absence of effect
- Effect Size
  - In simple terms this measures the strength of the observed effect, so a large effect size indicates a (statistically) stronger result than a weak one
  - Calculated differently for different statistical models so you’ll have to find the appropriate one for your analysis method

Presenting you results

Extract as much as you can from your data

- You do need stats
- Check for interaction effects
- Where possible report statistical significance, observed power and effects sizes

Provide meaningful plots

- Always include
  - A measure of the variance (eg: 95% conf. int.)
  - Titles
  - Axis labels
  - Legends

Don't conclude more than what your analysis reveals or your test was setup to do

Further reading

Perceptual Audio Evaluation – Bech and Zacharov

Journal of the Audio Engineering Society

- Rumsey, Toole

Journal of the Acoustical Society of America

(Journal of)Hearing Research – Elsevier

http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage28.htm - Chi Square Tests